鐵達尼號生存預測

（Kaggle 競賽）

10636009　陳仙姁

10636007　謝旻儒

10636015　徐亦萱

目錄

一、摘要

鐵達尼號生存預測是個很有趣的二元分類問題，必須依據乘客僅限的資訊（包括乘客的性別、姓名、出發港口、住艙等級、房間號碼、年齡、船上兄弟姊妹及配偶的數量、船上父母及小孩的數量、票價、票號這些特徵，使用訓練資料集去訓練出預測模型，分析什麼類型的人更可能在鐵達尼號沈船的意外中生存下來。

我們使用了三個機器學習的演算法來預測乘客的死活，分別是 SVM、KNN 及 Decision-Tree，再結合整體學習的投票法，來讓預測結果更加準確。

二、介紹

為了避免再度發生鐵達尼號沈船這樣的悲劇，我們想要提升乘客自身的存活率，所以使用機器學習的演算法來進行分析，找出具有哪些特徵的乘客較容易生存，建議沒有具備此特徵的人，務必要再三考慮是否要搭乘郵輪，免得丟了寶貴性命。

三、資料集介紹(含資料特徵)及資料集來源

● 資料集介紹

欄位名稱中英對照			
Variable	**中文**	**Definition**	**Key**
PassengerId	乘客編號		
survival	存活	Survival	0 = No, 1 = Yes
pclass	社會經濟地位	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	性別	Sex	
Age	年齡	Age in years	

欄位名稱中英對照			
sibsp	兄弟姊妹+老婆丈夫數量	# of siblings / spouses aboard the Titanic	
parch	父母小孩的數量	# of parents / children aboard the Titanic	
ticket	票的號碼	Ticket number	
fare	票價	Passenger fare	
cabin	住的艙等	Cabin number	
embarked	出發港口	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

- **資料集來源**

Titanic : Machine Learning from Disaster | Kaggle

https://www.kaggle.com/c/titanic/data

四、資料預處理

性別 Sex

- 將性別欄位的資料 male 改為 1，female 改為 0。

年齡 Age

- 將遺失值補上年齡的平均值。

票價 Fare

- 將遺失值補上票價的平均值。

出發的港口 Embarked

- 將遺失值補上出現最頻繁的值（S 港口），使用 One-Hot-Encoding。

特徵工程

兄弟姊妹與配偶的數量 SibSp & 父母與小孩的數量 Parch

- 將兄弟姊妹與配偶的數量 SibSp & 父母與小孩的數量 Parch，合併成家庭大小 Family_size，作為新的特徵。

姓名 Name

- 將姓名中的稱謂分割出來，並將少數稱謂合併至人數較多的稱謂，其中 Other 為 Doctor 稱謂的人（可能是醫生或者博士，且有男有女，故另分一類），最後留下 'Mr', 'Mrs', 'Miss', 'Master', 'Other' 這幾個稱謂，使用 One-Hot-Encoding 後，作為新的特徵。

票號 Ticket

- 將票號中的英文取出，遺失值統一以"X"替代，使用 One-Hot-Encoding 後，作為新的特徵。

住艙 Cabin

- 取出住艙中的甲板代號，遺失值統一以"noCabin"取代，使用 One-Hot-Encoding，作為新的特徵。

年齡 Age * 社會經濟地位 PClass

- 將年齡與社會經濟地位的數值相乘，作為新的特徵。

五、機器學習或深度學習方法 (使用何種方法)

特徵選擇

- Pclass 社會經濟地位
- Sex 性別
- Age 年齡
- Family_size 家庭大小
- Fare 票價
- 出發的港口 Embarked（One-Hot-Encoding）
    - ebk_S 登船港口（S）
    - ebk_C 登船港口（C）
    - ebk_Q 登船港口（Q）
- Title 稱謂（One-Hot-Encoding）
    - Mr
    - Mrs
    - Miss
    - Master
    - Other
- Cabin 住艙（One-Hot-Encoding）
    - cb_noCabin
    - cb_C
    - cb_E
    - cb_G
    - cb_D
    - cb_A
    - cb_B
    - cb_F
    - cb_T
- Age*PClass 年齡 Age * 社會經濟地位 PClass

演算法

- SVM
- KNN
- Decision Tree
- Ensemble Learning（Voting）

六、研究結果及討論 (含模型評估與改善)

Kaggle 成績截圖（Decision Tree）

Submission and Description	Public Score
submit.zip 16 hours ago by Joy Xie Ensemble Learning (Voting)	0.78468
Desktop.zip 16 hours ago by Yi-Xuan decision_tree	0.78947
submit.zip 2 days ago by Shiny Chen KNN	0.61244
submit.zip a month ago by Joy Xie SVM	0.77990

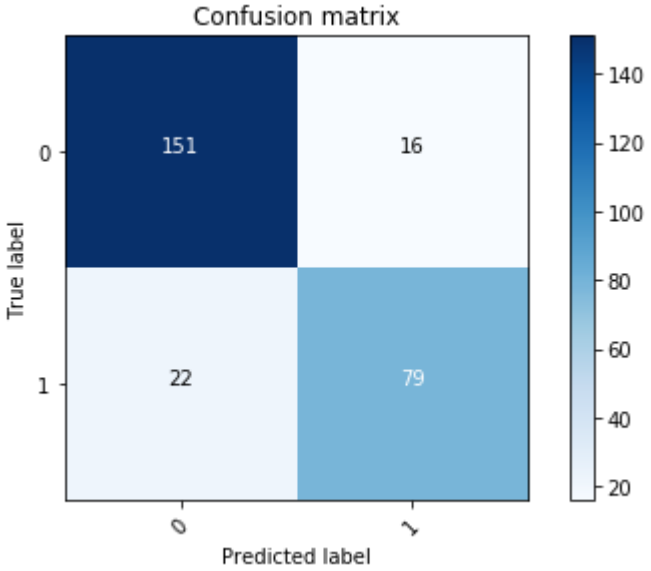
評估模型

模型	參數	訓練準確度	測試準確度	Kaggle
SVM	SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=3, gamma='auto', kernel='linear', max_iter=-1, probability=False,	0.8475120385232745	0.7873134328358209	0.77990

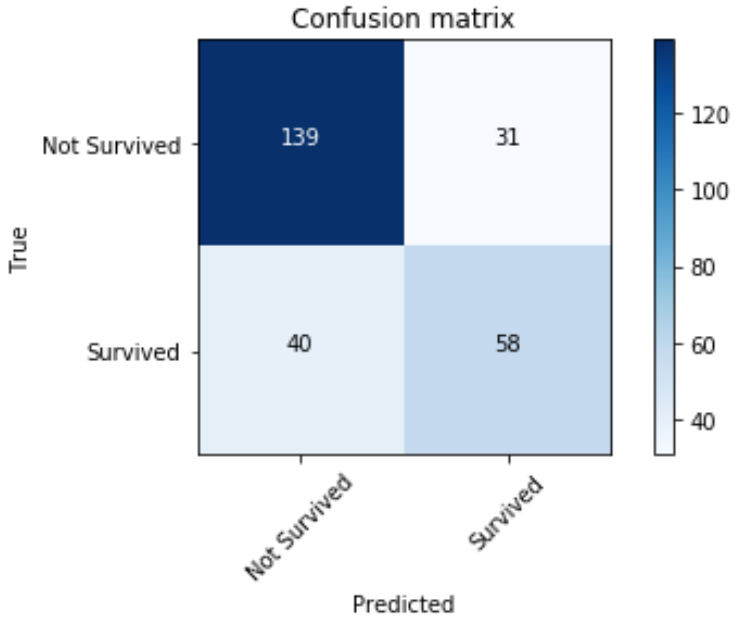
模型	參數	訓練準確度	測試準確度	Kaggle
	random_state=0, shrinking=True, tol=0.001, verbose=False)			
KNN	KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski', metric_params=None, n_jobs=1, n_neighbors=3, p=2, weights='uniform')	0.8218298555377207	0.7350746268656716	0.61244
Decision-Tree	DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=3, max_features=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, presort=False, random_state=0, splitter='best')	0.8282504012841091	0.8470149253731343	0.78947
Ensemble Learning (Voting)				0.78468

混淆矩陣

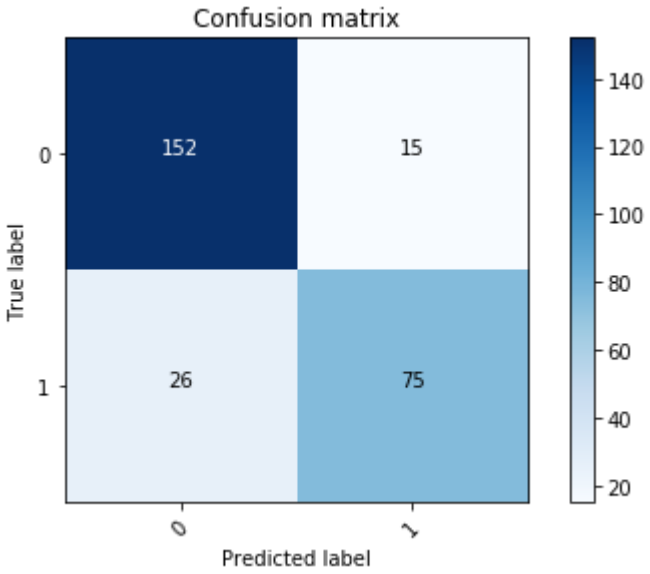
- SVM



● KNN



● Decision Tree



七、結論

我們有在網路上搜尋到使用隨機森林的預測方法，但是我們想要研究其他演算法預測此問題的準確度，因此選用 SVM、KNN、Decision Tree，最後加上 Ensemble Learning 的投票法（票票等值），得到的預測結果為 0.78468，比最好的分類器 Decision Tree 預測的準確率少了 0.00479，因此，我們從使用的演算法中判斷 Decision Tree 是最適合這個主題的分類器，若結合隨機森林會更好。

最好的演算法分類器準確率排序：

1. Decision Tree：0.78947
2. Ensemble Learning（Voting）：0.78468
3. SVM：0.77990
4. KNN：0.61244

八、參考文獻

- Basic Feature Engineering with the Titanic Data « triangleinequality
- [資料分析&機器學習] 第 4.1 講 : Kaggle 競賽-鐵達尼號生存預測(前 16%排名)
- 對 pandas 進行資料預處理的例項講解
- [資料分析&機器學習] 第 2.4 講：資料前處理(Missing data, One-hot encoding, Feature Scaling)
- [資料分析&機器學習] 第 3.5 講 : 決策樹(Decision Tree)以及隨機森林(Random Forest)介紹
- 机器学习（二）如何做到 Kaggle 排名前 2%
- kaggle 泰坦尼克号生存预测——六种算法模型实现与比较